Федеральное государственное автономное образовательное
учреждение высшего образования
«**СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ**»
Институт фундаментальной биологии и биотехнологий
Кафедра биофизики

УТВЕРЖДАЮ
Заведующий кафедрой
_____ / Кратасюк В.А.
«___» _____2016 г.

БАКАЛАВРСКАЯ РАБОТА

06.03.01. Биология

ИЗМЕНЧИВОСТЬ ГЕНОМОВ ПО СТЕПЕНИ НАРУШЕНИЯ ВТОРОГО ПРАВИЛА ЧАРГАФФА

Руководитель _____ д.ф.-м.н. М. Г. Садовский

Выпускник _____ Я. В. Петрова

Красноярск 2016

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Наука не стоит на месте, еще совсем недавно только начинали секвенировать геномы простейших организмов (*Haemophilus influenzae* 1995г.) [1], а на сегодняшний день уже очень много полностью аннотированных геномов различных организмов (включая человека), они нуждаются в обработке, но работать с ними неудобно из-за различной длины геномов.

Поэтому придумали простой способ сократить разницу между геномами посредством перехода от полного текста к его частотному словарю. Это простой, но продуктивный прием, позволяющий единообразно работать с текстами различной длины, сравнивать их, производить информационный анализ.

Цель моей работы заключалась в выявлении наличия зависимости между таксономическими признаками и степенью нарушения второго правила Чаргаффа в геноме.

Были поставлены следующие задачи:

1. Составить базу данных геномов бактерий, архей, эукариот.
2. Сравнить, совпадают ли разделения на группы по коэффициентам невязки с таксономией.

1. ОБЗОР ЛИТЕРАТУРЫ

1.1. Строение ДНК

С химической точки зрения ДНК — это длинная полимерная молекула, состоящая из повторяющихся блоков — нуклеотидов. Каждый нуклеотид состоит из азотистого основания, сахара (дезоксирибозы) и фосфатной группы. Связи между нуклеотидами в цепи образуются за счёт дезоксирибозы и фосфатной группы (фосфодиэфирные связи). В подавляющем большинстве случаев (кроме некоторых вирусов, содержащих одноцепочечную ДНК) макромолекула ДНК состоит из двух цепей, ориентированных азотистыми основаниями друг к другу. Эта двухцепочечная молекула спирализована. В целом структура молекулы ДНК получила название «двойной спирали».

В ДНК встречается четыре вида азотистых оснований (аденин, гуанин, тимин и цитозин). Азотистые основания одной из цепей соединены с азотистыми основаниями другой цепи водородными связями согласно принципу комплементарности: аденин соединяется только с тимином, гуанин — только с цитозином.

Последовательность нуклеотидов позволяет «кодировать» информацию о различных типах РНК, наиболее важными из которых являются информационные, или матричные (мРНК), рибосомальные (рРНК) и транспортные (тРНК). Все эти типы РНК синтезируются на матрице ДНК за счёт копирования последовательности ДНК в последовательность РНК, синтезируемой в процессе транскрипции, и принимают участие в биосинтезе белков (процессе трансляции). Помимо кодирующих последовательностей, ДНК клеток содержит последовательности, выполняющие регуляторные и структурные функции. Кроме того, в геноме эукариот часто встречаются участки, принадлежащие «генетическим паразитам», например, транспозонам [2].

1.2. Длина генома

Термин «геном» был предложен Г. Винклером в 1920 г. для описания совокупности генов, заключенных в гаплоидном наборе хромосом одного

биологического вида. Уже в то время подчеркивалось, что понятие генома в отличие от генотипа является характеристикой вида в целом, а не отдельной особи. Интенсивное исследование геномов
в течение последних пятидесяти лет заметно изменило представления об их строении и функционировании. С установлением полной первичной структуры генома человека, а также целого ряда животных, растений и микроорганизмов человечество вступило в «постгеномную эру». Сегодня термином «геном» обозначают совокупность ДНК гаплоидного набора хромосом, которые заключены в отдельной клетке зародышевой линии многоклеточного организма[3].

По соотношению размера генома и числа генов геномы могут быть разделены на два чётко выделенных класса:

1. Небольшие компактные геномы размером, как правило, не более 10 млн пар оснований, со строгим соответствием между размером генома и числом генов. Такими геномами обладают все вирусы и прокариоты. У этих организмов плотность генов составляет от 0,5 до 2 генов на тысячу пар оснований, а между генами имеются очень короткие участки, занимающие 10-15 % длины генома. Межгенные участки в таких геномах состоят главным образом из регуляторных элементов. Помимо вирусов и прокариот к этому классу могут быть отнесены и геномы большинства одноклеточных эукариот, хотя их геномы демонстрируют несколько меньшую зависимость между размером генома и числом генов, а размер генома может достигать 20 млн пар оснований.
2. Обширные геномы размером более 100 млн пар оснований, у которых нет чёткой взаимосвязи между размером генома и числом генов. К этому классу относятся большие геномы многоклеточных эукариот и некоторых одноклеточных эукариот. В отличие от геномов первой группы большинство нуклеотидов в геномах этого класса относятся к последовательностям, которые не кодируют ни белков, ни РНК[4].

Для того, чтобы понять, насколько сильно может различаться длина генома у разных организмов, приведу пример самого маленького генома - бактерии *Carsonella ruddii*, длина генома составляет 159 662 пары оснований [5], самым длинным расшифрованным геномом обладает ладанная сосна *Pinus taeda*, длина генома составляет 22,18 миллиардов пар оснований [6].

1.3. Частотный словарь

Нуклеотидные последовательности мы рассматриваем как текст, т.е. как последовательность символов из четырехбуквенного алфавита A,C,G,T. Длина генетического текста — количество N нуклеотидов в нуклеотидной последовательности.

Предполагается, что клеточные механизмы, связанные со считыванием наследственной информации, оперируют с отдельными, весьма малыми, фрагментами ДНК. В соответствии с этим перейдем от рассмотрения молекулы как целого к изучению совокупности ее фрагментов фиксированной длины (слов).

Список всех слов длины q, входящих в данный текст, называется q-носителем данного текста. Если каждому слову q-носителя сопоставить частоту его встречаемости в тексте, получим частотный словарь длины q [7]. Мы будем рассматривать словари при q=2-6 нуклеотидов.

Частоты слов были рассчитаны по следующей формуле:

$$f_{\omega}=\frac{n_{\omega}}{N} \quad (1)$$

где n_{ω} — число копий каждого слова, ω – слово, N – общее число слов [8].

1.4. Правила Чаргаффа

Правила Чаргаффа — система эмпирически выявленных правил, описывающих количественные соотношения между различными типами азотистых оснований в ДНК. Были сформулированы в результате работы группы биохимика Эрвина Чаргаффа в 1949-1951 гг.

До работ группы Чаргаффа господствовала так называемая «тетрануклеотидная» теория, согласно которой ДНК состоит из повторяющихся блоков по четыре разных азотистых основания (аденин, тимин, гуанин и цитозин). Чаргаффу и сотрудникам удалось разделить нуклеотиды ДНК при помощи бумажной хроматографии и определить точные количественные соотношения нуклеотидов разных типов. Они

значительно отличались от эквимолярных, которых можно было бы ожидать, если бы все четыре основания были представлены в равных пропорциях.

В 1968 году Чаргафф выявил, что в каждой из нитей ДНК количество аденина приблизительно равно количеству тимина, а гуанина — цитозину: А~Т, Г~Ц. В 1990-х с развитием технологии секвенирования ДНК это правило было подтверждено.

Соотношения, выявленные Чаргаффом для аденина (А), тимина (Т), гуанина (Г) и цитозина (Ц), оказались следующими:

1. Количество аденина равно количеству тимина, а гуанина — цитозину: А=Т, Г=Ц.
2. Количество пуринов равно количеству пиримидинов: А+Г=Т+Ц.
3. Количество оснований с аминогруппами в положении 6 равно количеству оснований с кетогруппами в положении 6: А+Ц=Г+Т.

Вместе с тем, соотношение (А+Т):(Г+Ц) может быть различным у ДНК разных видов. У одних преобладают пары АТ, в других — ГЦ. Эти соотношения применялись только к одноцепочечной ДНК.

Правила Чаргаффа, наряду с данными рентгеноструктурного анализа, сыграли решающую роль в расшифровке структуры ДНК Дж. Уотсоном и Фрэнсисом Криком.

Позднее, когда было секвенировано больше геномов, стало понятно, что соотношение (А+Т):(Г+Ц) выполняется и для обеих нитей ДНК одновременно. Это и было названо вторым правилом Чаргаффа[9,10].

Известно, что второе правило Чаргаффа нарушается у геномов разных организмов по-разному. Особенно часто нарушается оно в митохондриальных геномах [11].

2. МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

2.1. Материалы

Ниже приведены три таблицы с перечислением геномов, для которых определялась невязка нарушения второго правила Чаргаффа.

Геномы взяты из базы данных EMBL - http://www.embl.org/.

Таблица 1. Геномы архей, всего 109.

Отдел A1 *Crenarchaeota*	Отдел A2 *Euryarchaeota*
Acidianus hospitalis	*Archaeoglobus fulgidus*
Acidilobus saccharovorans	*Archaeoglobus profundus*
Aeropyrum camini	*Archaeoglobus sulfaticallidus*
Aeropyrum pernix	*Archaeoglobus veneficus*
Caldisphaera lagunensis	*Methanomassiliicoccus intestinalis*
Caldivirga maquilingensis	*Methanomethylophilus alvus*
Desulfurococcus fermentans	*Methanoplasma termitum*
Desulfurococcus mucosus	*Ferroglobus placidus*
Fervidicoccus fontis	*Ferroplasma acidarmanus*
Hyperthermus butylicus	*Geoglobus acetivorans*
Ignicoccus hospitalis	*Halalkalicoccus jeotgali*
Ignisphaera aggregans	*Haloarcula hispanica*
Metallosphaera cuprina	*Halobacterium salinarum*
Metallosphaera sedula	*Haloferax mediterranei*
Pyrobaculum arsenaticum	*Haloferax volcanii*
Pyrobaculum calidifontis	*Halogeometricum borinquense*
Pyrolobus fumarii	*Halomicrobium mukohataei*
Sulfolobus tokodaii	*Halopiger xanaduensis*
Thermogladius cellulolyticus	*Haloquadratum walsbyi*
Thermoproteus tenax	*Halorhabdus tiamatea*
Vulcanisaeta distributa	*Halorubrum lacusprofundi*
Vulcanisaeta moutnovskia	*Halostagnicola larsenii*
	Halovivax ruber
Отдел A3 *Korarchaeota*	*Methanobacterium formicicum*
Korarchaeum cryptofilum	*Methanobacterium lacus*

	Methanobacterium paludis
Отдел A4 Nanoarchaeota	*Methanobrevibacter ruminantium*
Nanoarchaeum equitans	*Methanobrevibacter smithii*
	Methanocaldococcus fervens
Без отдела	*Methanocaldococcus infernus*
Nitrosopumilus koreensis	*Methanocaldococcus jannaschii*
Nitrososphaera evergladensis	*Methanocaldococcus vulcanius*
Cenarchaeum symbiosum	*Methanocella arvoryzae*
Nitrosopumilus maritimus	*Methanocella conradii*
	Methanocella paludicola
	Methanococcoides burtonii
	Methanococcoides methylutens
	Methanococcus aeolicus
	Methanococcus vannielii
	Methanococcus voltae
	Methanocorpusculum labreanum
	Methanoculleus bourgensis
	Methanoculleus marisnigri
	Methanohalobium evestigatum
	Methanohalophilus mahii
	Methanolacinia petrolearia
	Methanolobus psychrophilus
	Methanomethylovorans hollandica
	Methanopyrus kandleri
	Methanoregula boonei
	Methanoregula formicica
	Methanosaeta concilii

	Methanosaeta harundinacea
	Methanosalsum zhilinae
	Methanosarcina acetivorans
	Methanosarcina thermophila
	Methanosarcina vacuolata
	Methanosphaera stadtmanae
	Methanosphaerula palustris
	Methanospirillum hungatei
	Methanothermobacter marburgensis
	Methanothermobacter thermautotrophicus
	Methanothermococcus okinawensis
	Methanothermus fervidus
	Methanotorris igneus
	Natrialba magadii
	Natrinema pellirubrum
	Natronobacterium gregoryi
	Natronococcus occultus
	Natronomonas moolapensis
	Natronomonas pharaonis
	Picrophilus torridus
	Pyrococcus abyssi
	Pyrococcus furiosus
	Pyrococcus horikoshii
	Thermococcus barophilus
	Thermococcus kodakarensis
	Thermococcus litoralis
	Thermococcus nautili

	Thermoplasma acidophilum
	Thermoplasma volcanium

Таблица 2. Геномы бактерий, всего 240.

Без отдела	Отдел B12 *Proteobacteria*
Aminobacterium colombiense	*Acetobacter pasteurianus*
Anaerobaculum mobile	*Achromobacter xylosoxidans*
Candidatus Koribacter versatilis	*Acidovorax avenae*
Marivirga tractuosa	*Acidovorax citrulli*
Oceanithermus profundus	*Acinetobacter calcoaceticus*
Pseudopedobacter saltans	*Acinetobacter oleivorans*
	Actinobacillus pleuropneumoniae
Отдел B1 *Aquificae*	*Actinobacillus succinogenes*
Aquifex aeolicus	*Actinobacillus suis*
	Advenella kashmirensis
Отдел B2 *Thermotogae*	*Advenella mimigardefordensis*
Fervidobacterium pennivorans	*Aeromonas media*
Pseudothermotoga thermarum	*Aeromonas veronii*
	Aggregatibacter aphrophilus
Отдел B4 *Deinococcus-Thermus*	*Alcanivorax borkumensis*
Deinococcus maricopensis	*Alcanivorax dieselolei*
Marinithermus hydrothermalis	*Alcanivorax pacificus*
Meiothermus ruber	*Alicycliphilus denitrificans*
Meiothermus silvanus	*Allochromatium vinosum*
	Arcobacter nitrofigilis
Отдел B6 *Chloroflexi*	*Arthrobacter chlorophenolicus*

Chloroflexus aurantiacus	*Azorhizobium caulinodans*
Herpetosiphon aurantiacus	*Azotobacter chroococcum*
Roseiflexus castenholzii	*Bacillus anthracis*
	Bacillus cellulosilyticus
Отдел B8 *Nitrospira*	*Bacillus cereus*
Desulfovibrio salexigens	*Bacillus methanolicus*
Leptospirillum ferriphilum	*Bacillus subtilis*
	Bacillus thuringiensis
Отдел B9 *Deferribacteres*	*Bacteriovorax marinus*
Denitrovibrio acetiphilus	*Bdellovibrio exovorus*
Flexistipes sinusarabici	*Bordetella avium*
	Bordetella holmesii
Отдел B10 *Cyanobacteria*	*Campylobacter coli*
Acaryochloris marina	*Campylobacter curvus*
Anabaena cylindrica	*Campylobacter volucris*
Chroococcidiopsis thermalis	*Candidatus Kinetoplastibacterium galatii*
Dactylococcopsis salina	*Candidatus Portiera aleyrodidarum*
	Candidatus Symbiobacter mobilis
Отдел B11 *Chlorobi*	*Chromohalobacter salexigens*
Pelodictyon luteolum	*Citrobacter freundii*
	Coxiella burnetii
Отдел B13 *Firmicutes*	*Desulfobacca acetoxidans*
Acetobacterium woodii	*Desulfohalobium retbaense*
Acetohalobium arabaticum	*Desulfurobacterium thermolithotrophum*
Acholeplasma laidlawii	*Erwinia pyrifoliae*
Acholeplasma oculi	*Escherichia coli*
Acidaminococcus fermentans	*Ferrimonas balearica*

Ammonifex degensii	*Geobacter metallireducens*
Anaerococcus prevotii	*Gluconobacter oxydans*
Anoxybacillus flavithermus	*Haemophilus ducreyi*
Bacteroides salanitronis	*Haemophilus influenzae*
Bacteroides vulgatus	*Haliangium ochraceum*
Brevibacillus brevis	*Halomonas elongata*
Brevibacillus laterosporus	*Helicobacter cetorum*
Caldicellulosiruptor hydrothermalis	*Helicobacter mustelae*
Caldicellulosiruptor kristjanssonii	*Helicobacter pylori*
Clostridium botulinum	*Hippea maritima*
Dehalobacter restrictus	*Kangiella koreensis*
Desulfitobacterium hafniense	*Legionella pneumophila*
Desulfosporosinus meridiei	*Leptothrix cholodnii*
Desulfotomaculum acetoxidans	*Mannheimia haemolytica*
Desulfotomaculum gibsoniae	*Myxococcus stipitatus*
Desulfotomaculum ruminis	*Neisseria gonorrhoeae*
Exiguobacterium sibiricum	*Neisseria meningitidis*
Geobacillus kaustophilus	*Nitratifractor salsuginis*
Geobacillus thermodenitrificans	*Nitrosomonas eutropha*
Halanaerobium praevalens	*Novosphingobium aromaticivorans*
Halobacillus halophilus	*Paracoccus aminophilus*
Halobacteroides halobius	*Pectobacterium wasabiae*
Kyrpidia tusciae	*Phaeobacter gallaeciensis*
Lactobacillus casei	*Phaeobacter inhibens*
Lactobacillus helveticus	*Pseudomonas aeruginosa*
Lactobacillus reuteri	*Pseudomonas balearica*
Lactobacillus rhamnosus	*Pseudomonas fluorescens*

Lactococcus lactis	*Pseudomonas mosselii*
Listeria ivanovii	*Pseudomonas putida*
Listeria monocytogenes	*Ralstonia solanacearum*
Mycoplasma gallisepticum	*Rhizobium etli*
Mycoplasma hyorhinis	*Roseibacterium elongatum*
Mycoplasma pneumoniae	*Salmonella enterica*
Paenibacillus borealis	*Starkeya novella*
Paenibacillus durus	*Xylella fastidiosa*
Paenibacillus mucilaginosus	*Yersinia pestis*
Paenibacillus polymyxa	*Yersinia pseudotuberculosis*
Paenibacillus sabinae	*Yersinia ruckeri*
Paenibacillus terrae	
Peptoclostridium difficile	Отдел B14 *Bacteroidetes*
Roseburia hominis	*Aequorivita sublithincola*
Staphylococcus aureus	*Alistipes finegoldii*
Staphylococcus pseudintermedius	*Barnesiella viscericola*
Streptococcus agalactiae	*Capnocytophaga canimorsus*
Streptococcus mutans	*Cellulophaga baltica*
Streptococcus pneumoniae	*Cellulophaga lytica*
Streptococcus pseudopneumoniae	*Dyadobacter fermentans*
Streptococcus pyogenes	*Emticicia oligotrophica*
Veillonella parvula	*Flexibacter litoralis*
Weissella ceti	*Fluviicola taffensis*
	Haliscomenobacter hydrossis
	Muricauda ruestringensis
Отдел B15 *Planctomycetes*	*Niabella soli*
Pirellula staleyi	*Odoribacter splanchnicus*

Planctopirus limnophilus	*Ornithobacterium rhinotracheale*
Rubinisphaera brasiliensis	*Owenweeksia hongkongensis*
	Pedobacter heparinus
Отдел B16 *Chlamydiae*	*Pelobacter carbinolicus*
Chlamydia avium	*Pelobacter propionicus*
Chlamydia pecorum	*Porphyromonas asaccharolytica*
	Rhodothermus marinus
Отдел B17 *Spirochaetes*	*Riemerella anatipestifer*
Borrelia anserina	*Salinibacter ruber*
Borrelia bissettii	*Weeksella virosa*
Borrelia coriaceae	
Borrelia hermsii	Отдел B24 *Actinobacteria*
Borrelia recurrentis	*Acidothermus cellulolyticus*
Brachyspira hyodysenteriae	*Actinoplanes missouriensis*
Treponema brennaborense	*Actinosynnema mirum*
Treponema caldaria	*Adlercreutzia equolifaciens*
	Amycolatopsis methanolica
Отдел B19 *Fusobacteria*	*Amycolatopsis orientalis*
Ilyobacter polytropus	*Atopobium parvulum*
	Bifidobacterium animalis
Отдел B21 *Verrucomicrobia*	*Bifidobacterium bifidum*
Akkermansia muciniphila	*Blastococcus saxobsidens*
Opitutus terrae	*Brachybacterium faecium*
	Corynebacterium glyciniphilum
Отдел B22 *Dictyoglomi*	*Corynebacterium jeikeium*
Dictyoglomus turgidum	*Eggerthella lenta*
	Geodermatophilus obscurus

	Intrasporangium calvum
	Isoptericola variabilis
	Jonesia denitrificans
	Kocuria rhizophila
	Kribbella flavida
	Kutzneria albida
	Kytococcus sedentarius
	Mycobacterium tuberculosis
	Nakamurella multipartita
	Nocardiopsis alba
	Parascardovia denticolens
	Rubrobacter xylanophilus
	Sanguibacter keddieii
	Streptomyces avermitilis
	Streptomyces cattleya
	Streptomyces fulvissimus

Таблица 3. Геномы эукариот, всего 100.

Anopheles gambiae str. PEST chromosome X	*Kazachstania naganishii CBS 8797 chromosome 4*
Arabidopsis thaliana chromosome 4, long arm	*Kluyveromyces lactis NRRL Y-1140 chromosome D*
Arabidopsis thaliana chromosome 4, short arm	*Kluyveromyces marxianus DMKU3-1042 DNA, chromosome 4*
Ashbya gossypii ATCC 10895 chromosome IV	*Komagataella pastoris CBS 7435 chromosome 4, complete replicon*
Aspergillus fumigatus Af293 chromosome 4	*Lachancea kluyveri NRRL Y-12651 chromosome D*

Aspergillus nidulans FGSC A4 chromosome IV	*Lachancea thermotolerans CBS 6340 chromosome D*
Aspergillus niger supercontig Sc2, chromosome map 4R	*Leishmania braziliensis chromosome 4*
Aspergillus niger supercontig Sc7, chromosome map 4L	*Leishmania donovani BPK282A1 chromosome 4*
Aspergillus niger supercontig Sc19, chromosome map 4ER	*Leishmania major strain Friedlin, chromosome 4*
Bos taurus chromosome 4	*Leishmania mexicana, chromosome 4*
Bos taurus chromosome X	*Macaca fascicularis chromosome 4*
Brachypodium distachyon strain Bd21 chromosome 4	*Macaca fascicularis chromosome X*
Caenorhabditis briggsae AF16 draft chromosome, chrIV	*Macaca mulatta chromosome 4*
Caenorhabditis briggsae AF16 draft chromosome, chrX	*Macaca mulatta chromosome X*
Caenorhabditis elegans chromosome IV	*Medicago truncatula chromosome 4*
Caenorhabditis elegans chromosome X	*Nasonia vitripennis chromosome 4*
Callithrix jacchus chromosome 4	*Naumovozyma castellii CBS 4309, chromosome 4*
Callithrix jacchus chromosome X	*Naumovozyma dairenensis CBS 421, chromosome 4*
Candida albicans WO-1 chromosome 4 genomic scaffold	*Neospora caninum Liverpool, chromosome IV*
Candida dubliniensis CD36 chromosome 4	*Ornithorhynchus anatinus chromosome 4*
Candida glabrata CBS 138 chromosome D	*Oryctolagus cuniculus chromosome 4*
Canis lupus familiaris chromosome 4	*Oryza sativa Indica Group chromosome 4*
Cryptococcus neoformans var. neoformans B-3501A chromosome 4	*Oryzias latipes chromosome 4*
Cryptococcus neoformans var. neoformans JEC21 chromosome 4	*Ostreococcus lucimarinus CCE9901 chromosome 4*

Cryptosporidium parvum Iowa II chromosome 4	*Ovis aries chromosome 4*
Cucumis melo genomic chromosome, chr_4	*Pan troglodytes chromosome 4*
Cyanidioschyzon merolae strain 10D chromosome 4	*Phaeodactylum tricornutum chromosome 4*
Danio rerio chromosome 4	*Plasmodium falciparum 3D7 chromosome 4*
Debaryomyces hansenii CBS767 chromosome D	*Plasmodium knowlesi strain H chromosome 4*
Dictyostelium discoideum AX4 chromosome 4	*Plasmodium vivax chromosome 4*
Drosophila melanogaster chromosome 4	*Pongo abelii chromosome 4*
Drosophila melanogaster chromosome X	*Rattus norvegicus chromosome 4*
Drosophila simulans chromosome 4	*Saccharomyces cerevisiae chromosome IV*
Drosophila simulans chromosome X	*Scheffersomyces stipitis CBS 6054 chromosome 4*
Drosophila yakuba strain Tai18E2 chromosome 4	*Schistosoma mansoni strain Puerto Rico chromosome 4*
Drosophila yakuba strain Tai18E2 chromosome X	*Solanum lycopersicum chromosome ch04*
Encephalitozoon cuniculi GB-M1 chromosome IV	*Solanum pennellii chromosome ch04*
Encephalitozoon intestinalis ATCC 50506 chromosome IV	*Sorghum bicolor chromosome 4*
Equus caballus chromosome 4	*Sus scrofa chromosome 4*
Equus caballus chromosome X	*Taeniopygia guttata chromosome 4*
Eremothecium cymbalariae DBVPG#7215 chromosome 4	*Tetrapisispora blattae CBS 6284 chromosome 4*
Fusarium graminearum chromosome 4	*Tetrapisispora phaffii CBS 4417 chromosome 4*
Fusarium oxysporum f. sp. lycopersici 4287 chromosome 4	*Thalassiosira pseudonana CCMP1335 chromosome 4*

Fusarium verticillioides 7600 chromosome 4	*Theileria annulata chromosome 4*
Gallus gallus chromosome 4	*Theileria orientalis strain Shintoku DNA, chromosome 4*
Glycine max chromosome 4	*Thielavia terrestris NRRL 8126 chromosome 4*
Gorilla gorilla gorilla assembly, supercontig chr4	*Torulaspora delbrueckii CBS 1146 chromosome 4*
Homo sapiens chromosome 4	*Trypanosoma brucei gambiense DAL972 chromosome 4*
Homo sapiens chromosome X	*Yarrowia lipolytica CLIB122 chromosome D*
Kazachstania africana CBS 2517 chromosome 4	*Zea mays chromosome 4*
	Zygosaccharomyces rouxii strain CBS732 chromosome D

2.2. Методы

Коэффициент невязки был найден с помощью специального программного обеспечения по следующей формуле:

$$M_q = \frac{\sqrt{\sum_{\omega} \left(f_{\omega} - f_{\acute{\omega}}\right)^2} * 1}{\|\Omega\|} \qquad (2)$$

где $\Omega = \frac{4^q}{2}$; f_{ω} – частота встречаемости слова; $f_{\acute{\omega}} - ¿$ частота встречаемости комплиментарного слова в другой нити ДНК; $\omega - ¿$ слово; $q - ¿$ длина слова.

Невязка измеряет меру отклонения внутри словаря, очень близка к Евклидовой метрике, но не является ею.

Если бы частоты слов были равными, то невязка была бы равна нулю и второе правило Чаргаффа выполнялось с абсолютной точностью [12].

3. РЕЗУЛЬТАТЫ

Ниже приведены графики, на которых сравниваются невязки разных видов организмов: архей, бактерий, эукариот.

Сравнивались значения невязок разных отделов при разной толщине словаря.

На графиках видно, что невязки родственных видов принимают близкие значения.

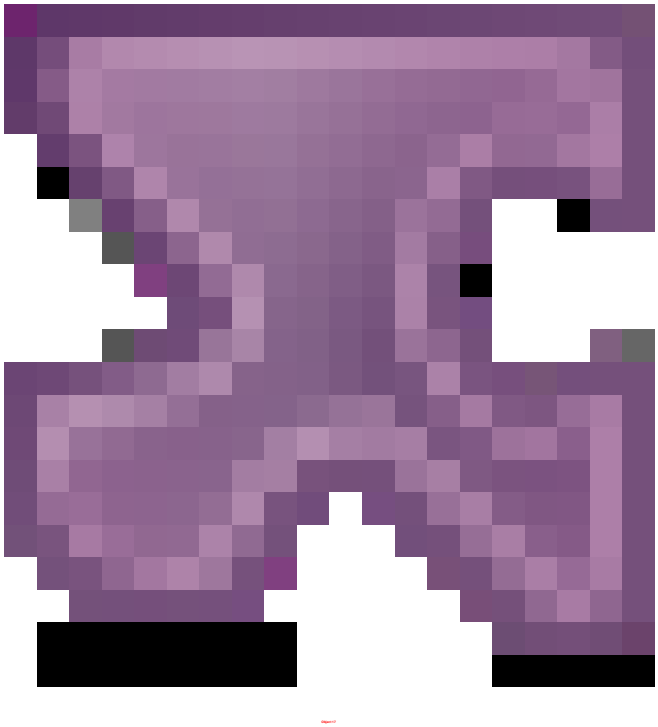

Рисунок 1 – Сравнение коэффициентов невязки для слов, длиной 1, разных геномов архей.

Рисунок 2 – Сравнение коэффициентов невязки для слов, длиной 2, разных геномов архей.

Рисунок 3 – Сравнение коэффициентов невязки для слов, длиной 3, разных геномов архей.

Рисунок 4 – Сравнение коэффициентов невязки для слов, длиной 4, разных геномов архей.

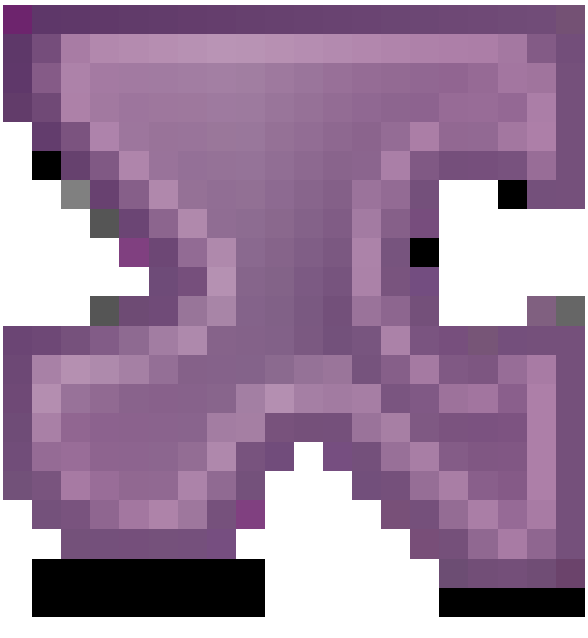

Рисунок 5 – Сравнение коэффициентов невязки для слов, длиной 5, разных геномов архей.

Рисунок 6 – Сравнение коэффициентов невязки для слов, длиной 6, разных геномов архей.

Рисунок 7 – Сравнение коэффициентов невязки для слов, длиной 1, разных геномов бактерий.

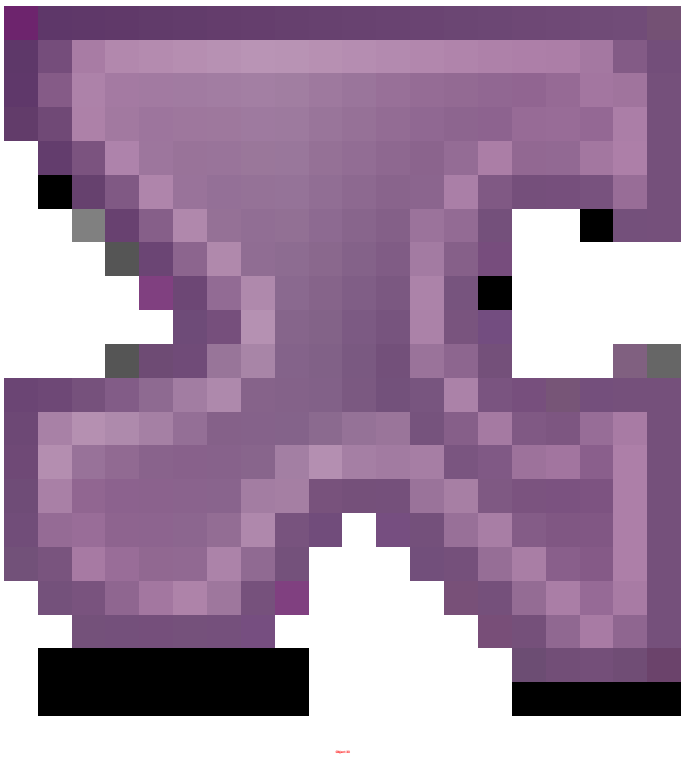

Рисунок 8 – Сравнение коэффициентов невязки для слов, длиной 2, разных геномов бактерий.

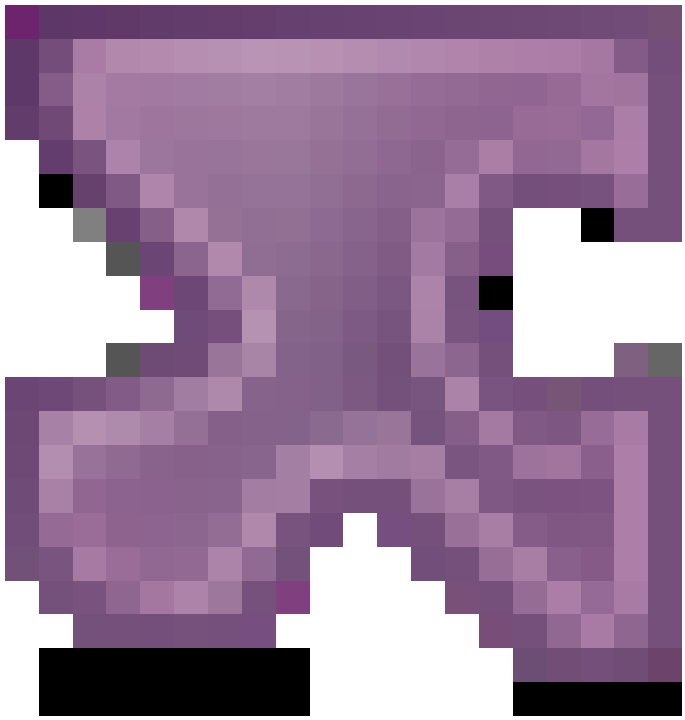

Рисунок 9 – Сравнение коэффициентов невязки для слов, длиной 3, разных геномов бактерий.

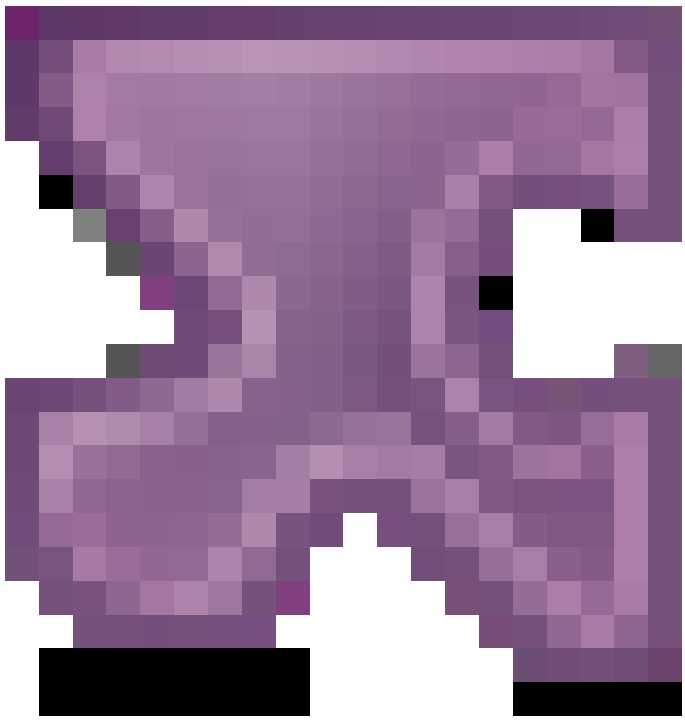

Рисунок 10 – Сравнение коэффициентов невязки для слов, длиной 4, разных геномов бактерий.

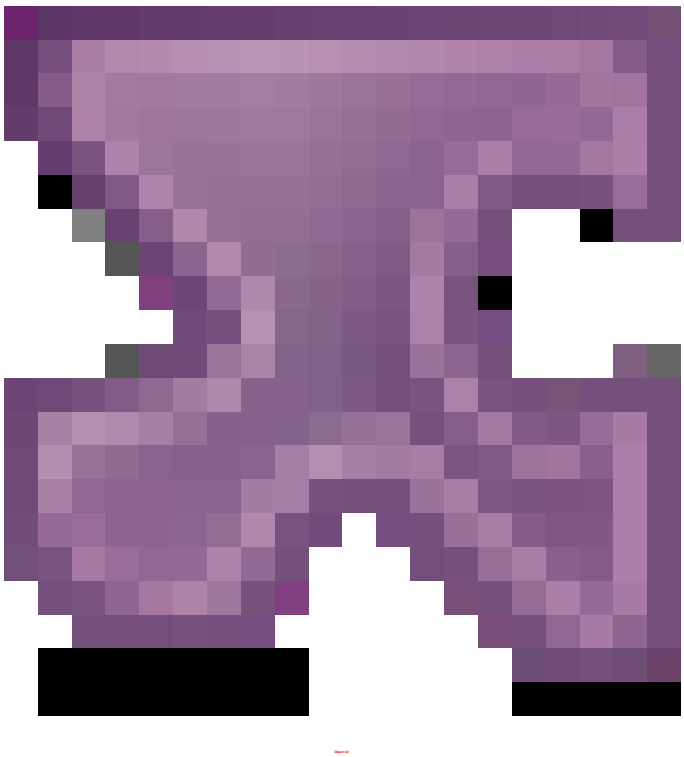

Рисунок 11 – Сравнение коэффициентов невязки для слов, длиной 5, разных геномов бактерий.

Рисунок 12 – Сравнение коэффициентов невязки для слов, длиной 6, разных геномов бактерий.

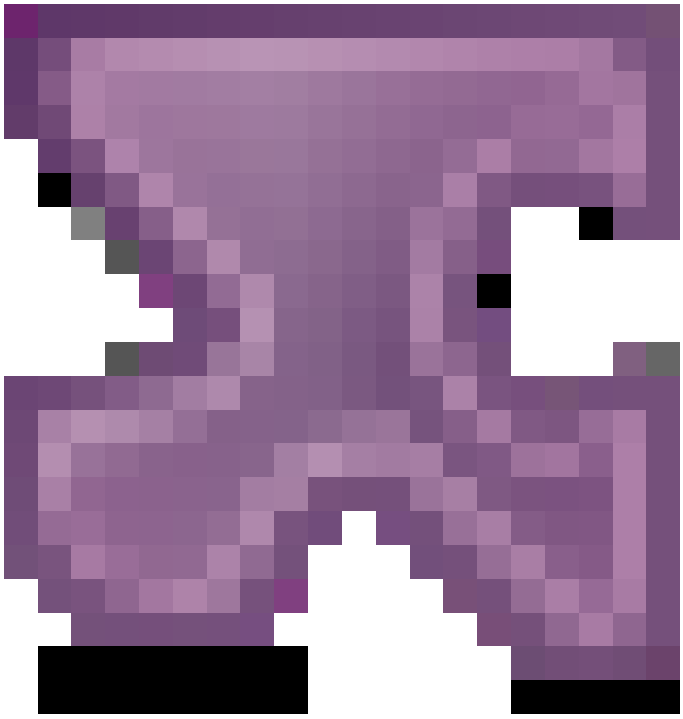

Рисунок 13 – Сравнение коэффициентов невязки для слов, длиной 1, разных геномов эукариот.

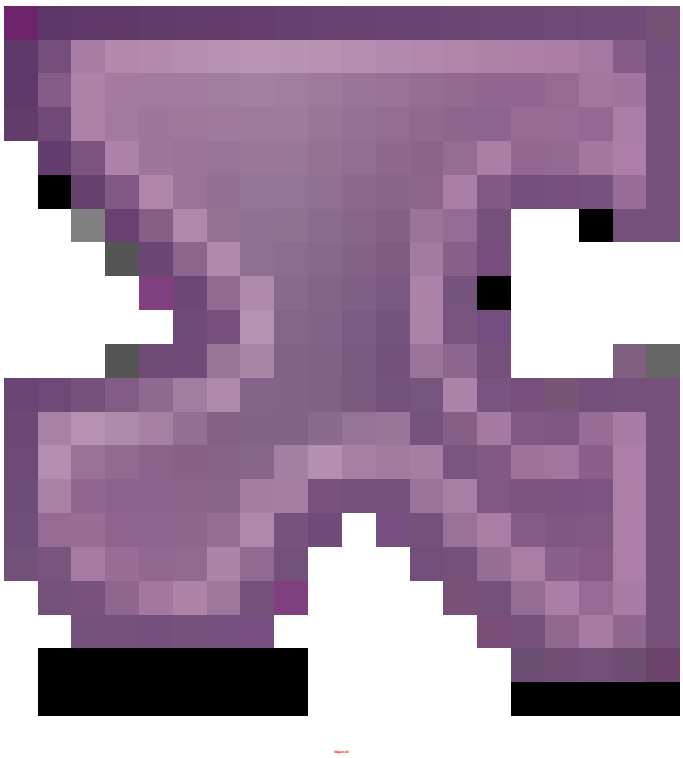

Рисунок 14 – Сравнение коэффициентов невязки для слов, длиной 2, разных геномов эукариот.

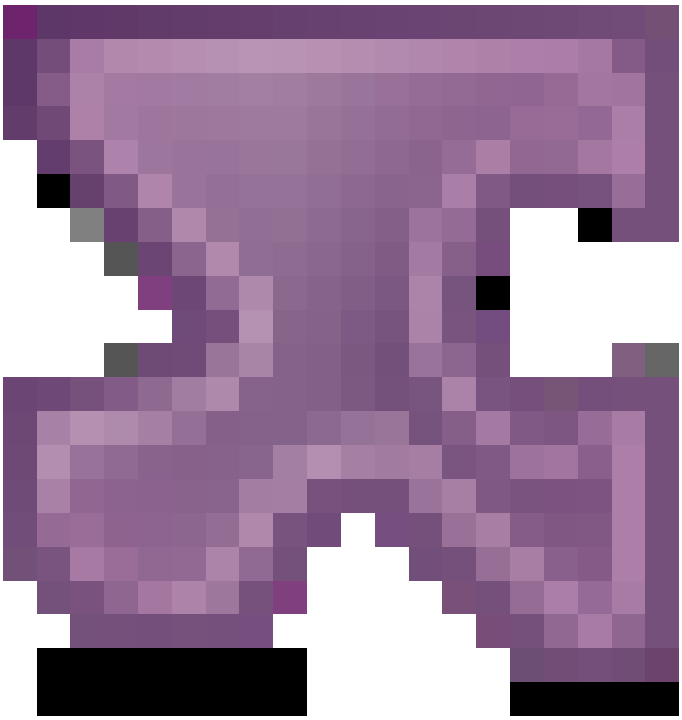

Рисунок 15 – Сравнение коэффициентов невязки для слов, длиной 3, разных геномов эукариот.

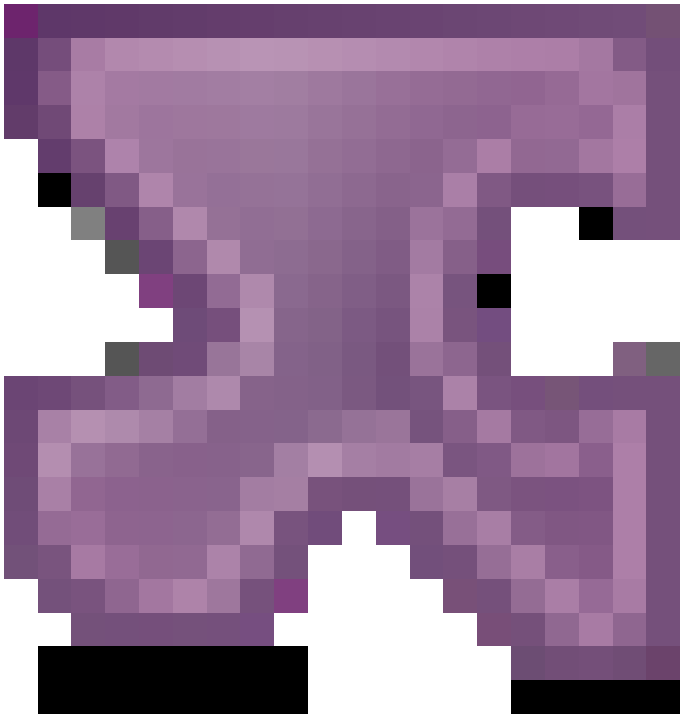

Рисунок 16 – Сравнение коэффициентов невязки для слов, длиной 4, разных геномов эукариот.

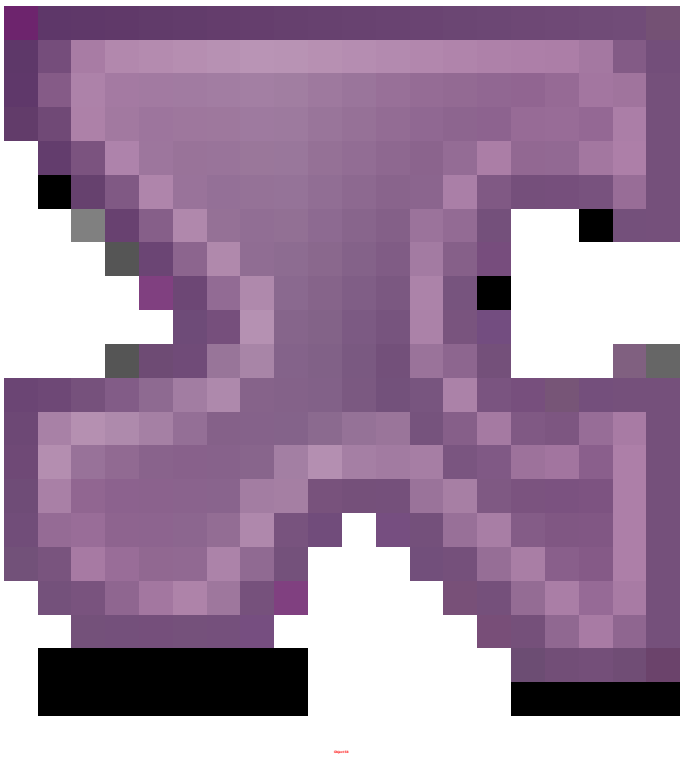

Рисунок 17 – Сравнение коэффициентов невязки для слов, длиной 5, разных геномов эукариот.

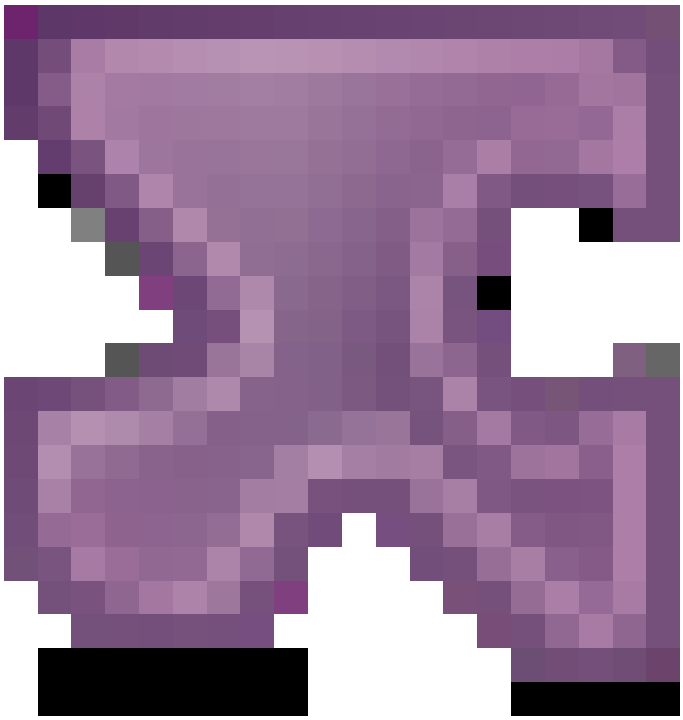

Рисунок 18 – Сравнение коэффициентов невязки для слов, длиной 6, разных геномов эукариот.

Ниже приведены графики, на которых сравниваются средние значения невязок разных видов организмов: архей, бактерий, эукариот; отделы разделены на порядки по таксономии.

На графиках видно, что значение невязки падает экспоненциально с ростом толщины словаря, это верно для всех организмов: архей, бактерий и эукариот.

Рисунок 19 – Среднее значение невязки отдела *Crenarchaeota*, разбитого на порядки.

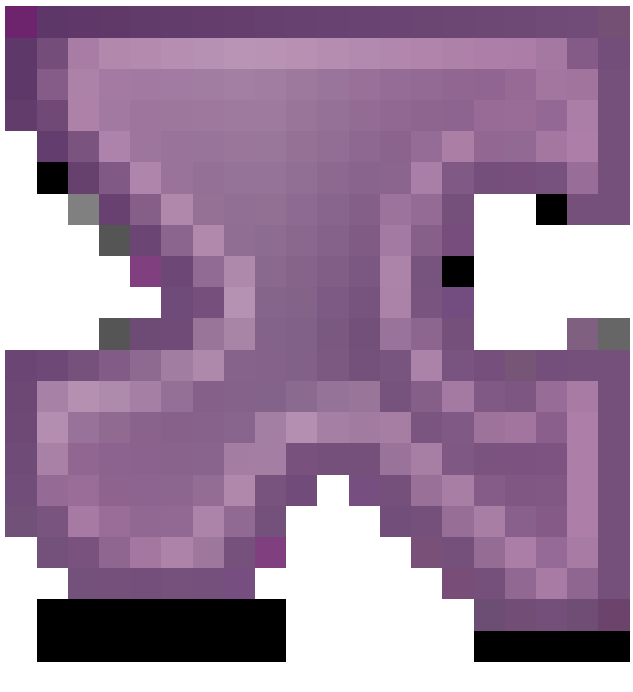

Рисунок 20 – Среднее значение невязки отдела *Euryarchaeota*, разбитого на порядки.

Рисунок 21 – Значение невязки отдела *Korarchaeota*, вид *Korarchaeum cryptofilum*.

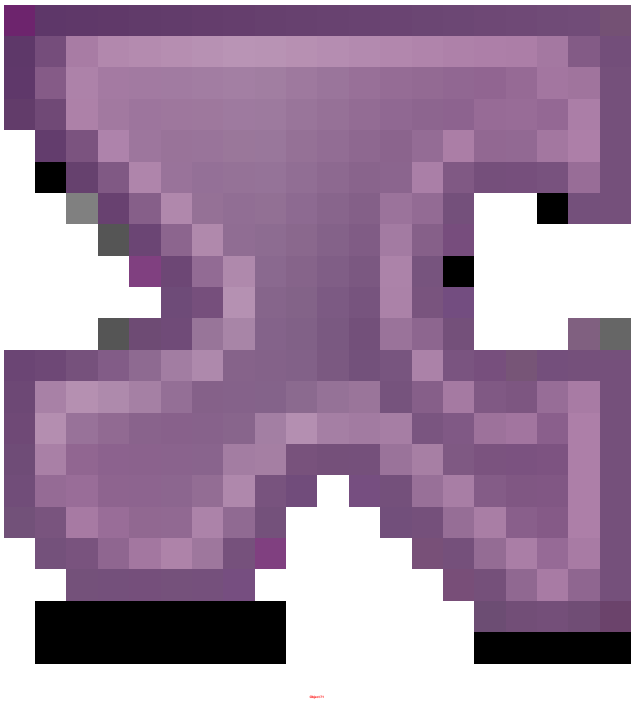

Рисунок 22 – Значение невязки отдела *Nanoarchaeota*, вид *Nanoarchaeum equitans*.

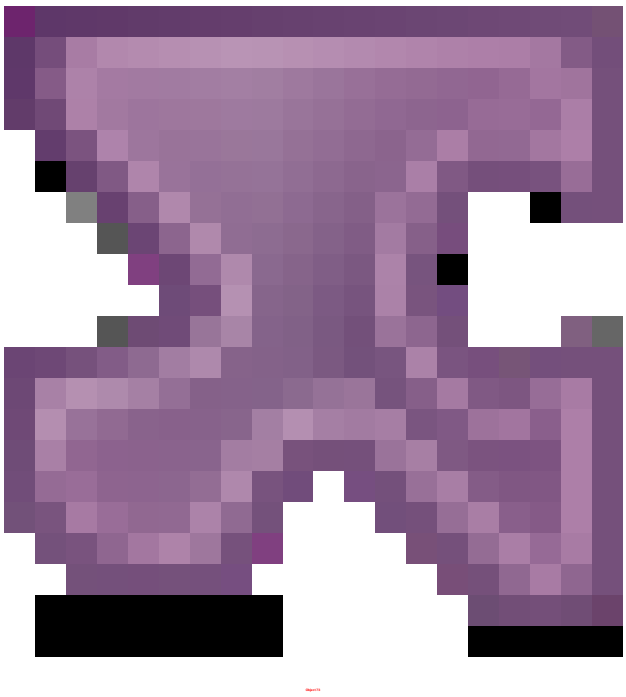

Рисунок 23 – Среднее значение невязки отдела *Thaumarchaeota*.

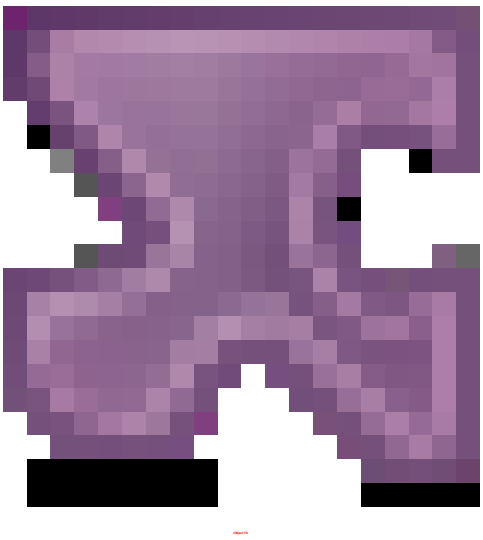

Рисунок 24 – Значение невязки отдела *Aquificae*, вид *Aquifex aeolicus*.

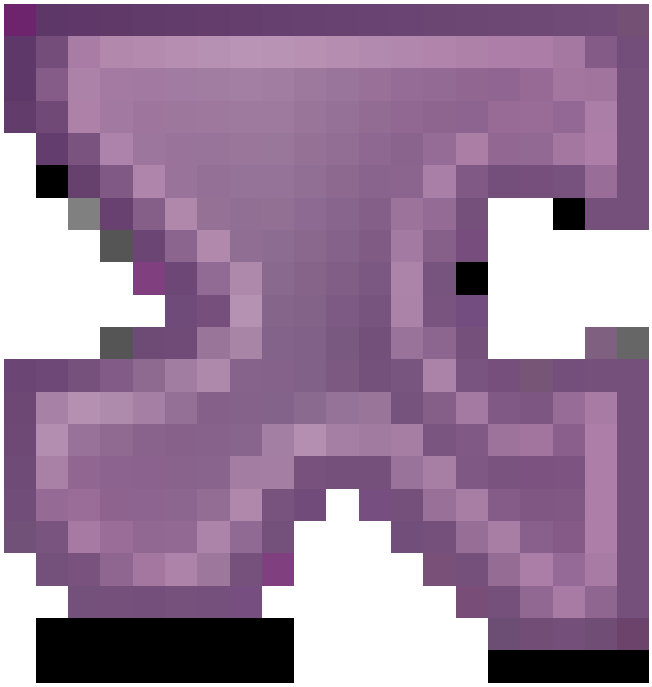

Рисунок 25 – Среднее значение невязки отдела *Thermotogae*, порядок *Thermotogales*.

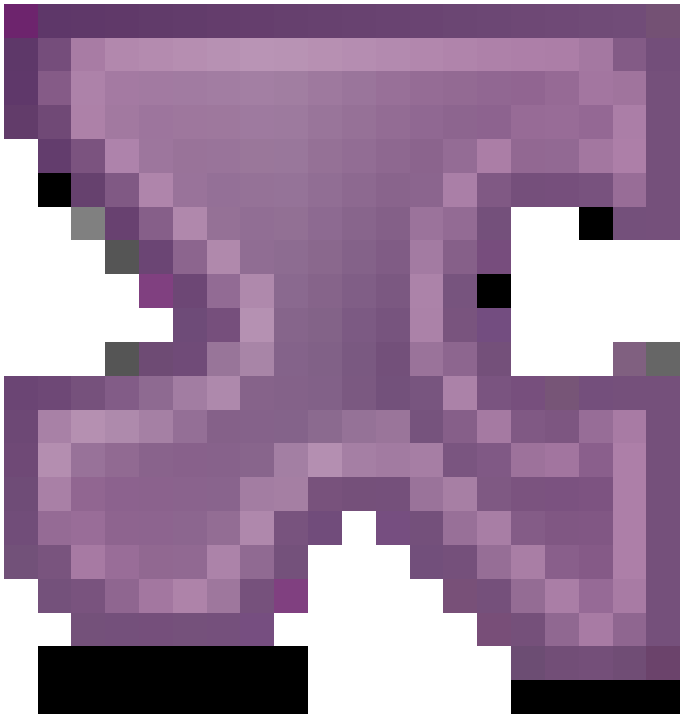

Рисунок 26 – Среднее значение невязки отдела *Deinococcus-Thermus*.

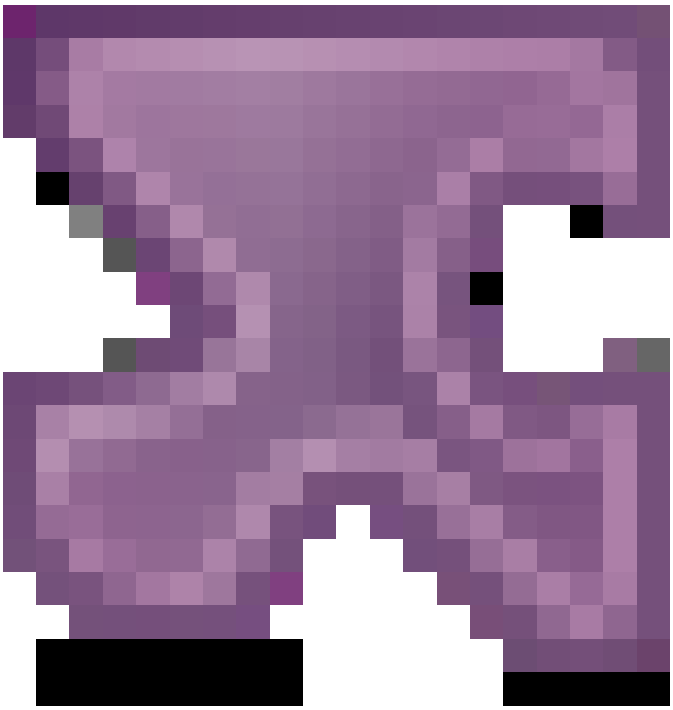

Рисунок 27 – Среднее значение невязки отдела *Chloroflexi*.

Рисунок 28 – Среднее значение невязки отдела *Nitrospira*.

Рисунок 29 – Среднее значение невязки отдела *Deferribacteres*, порядок *Deferribacterales*.

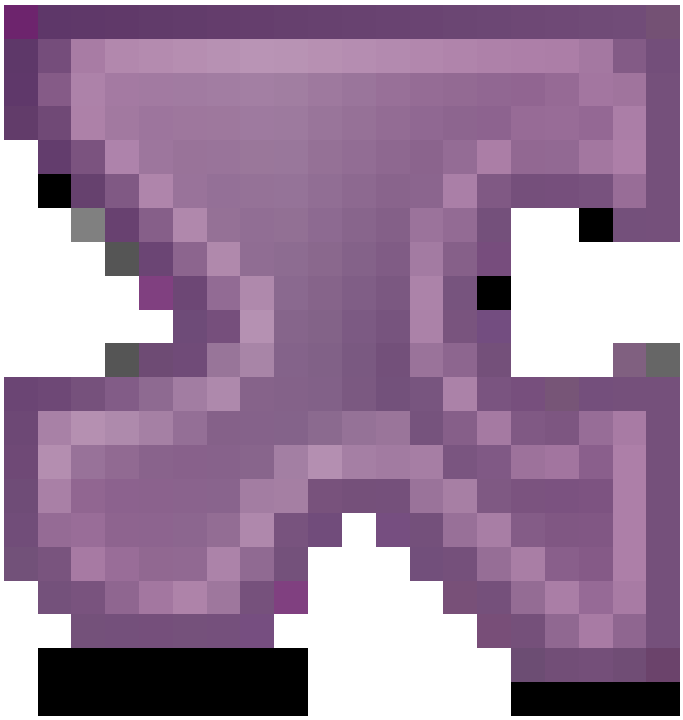

Рисунок 30 – Среднее значение невязки отдела *Cyanobacteria*.

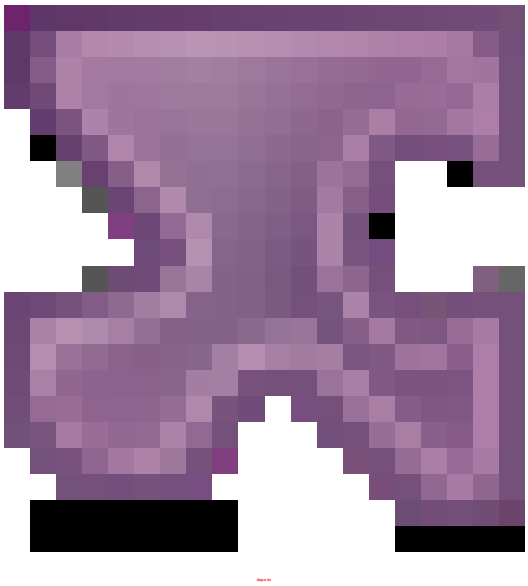

Рисунок 31 – Значение невязки отдела *Chlorobi*, вид *Pelodictyon luteolum*.

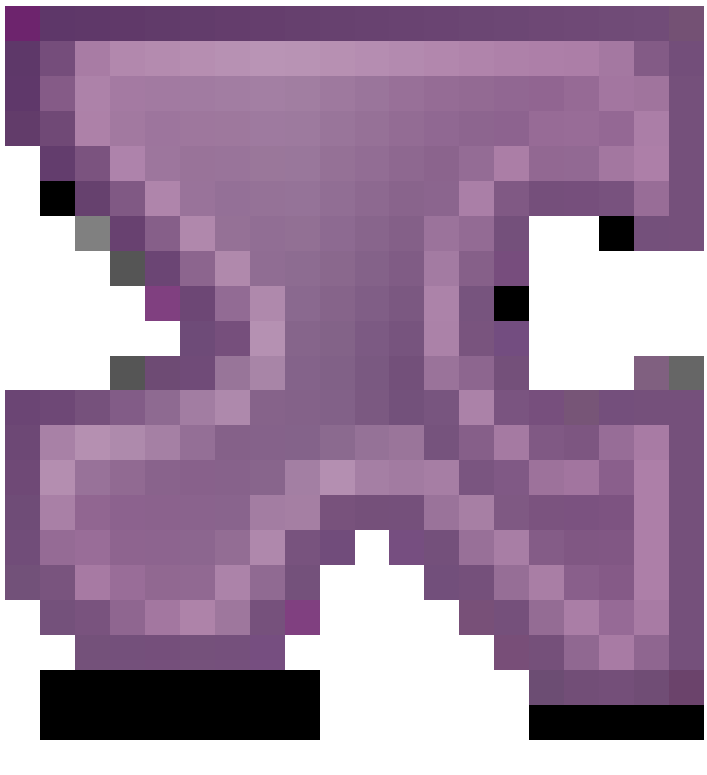

Рисунок 32 – Среднее значение невязки отдела *Proteobacteria*.

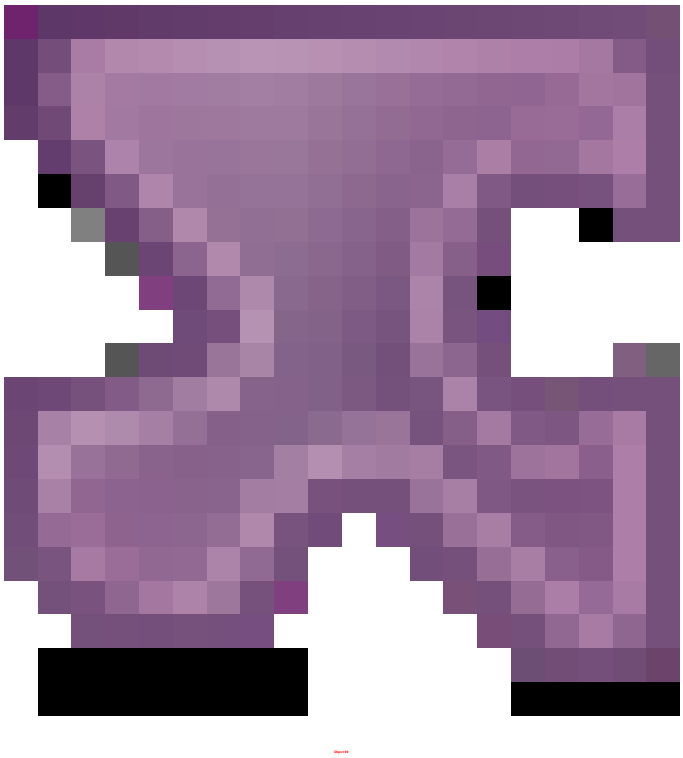

Рисунок 33 – Среднее значение невязки отдела *Firmicutes*.

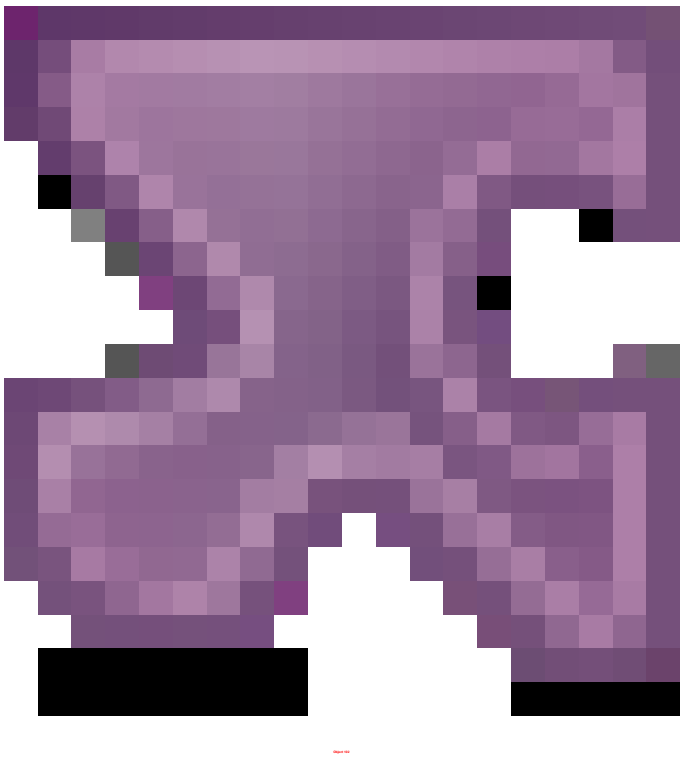

Рисунок 34 – Среднее значение невязки отдела *Bacteroidetes*.

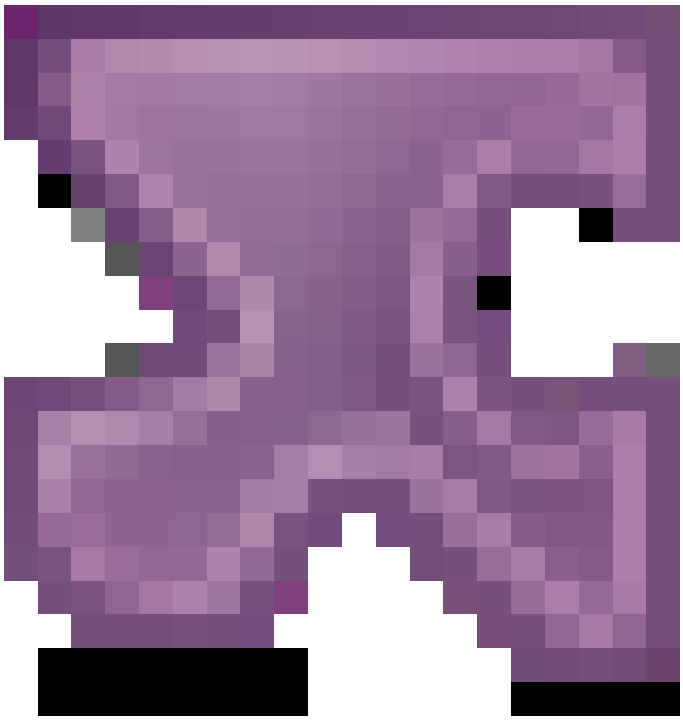

Рисунок 35 – Среднее значение невязки отдела *Planctomycetes*, порядок *Planctomycetales*.

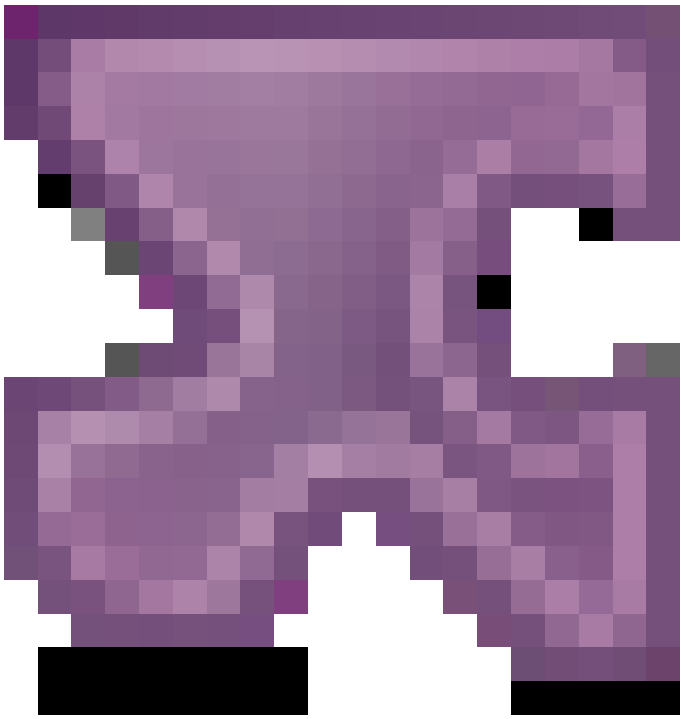

Рисунок 36 – Среднее значение невязки отдела *Chlamydiae*, порядок *Chlamydiales*.

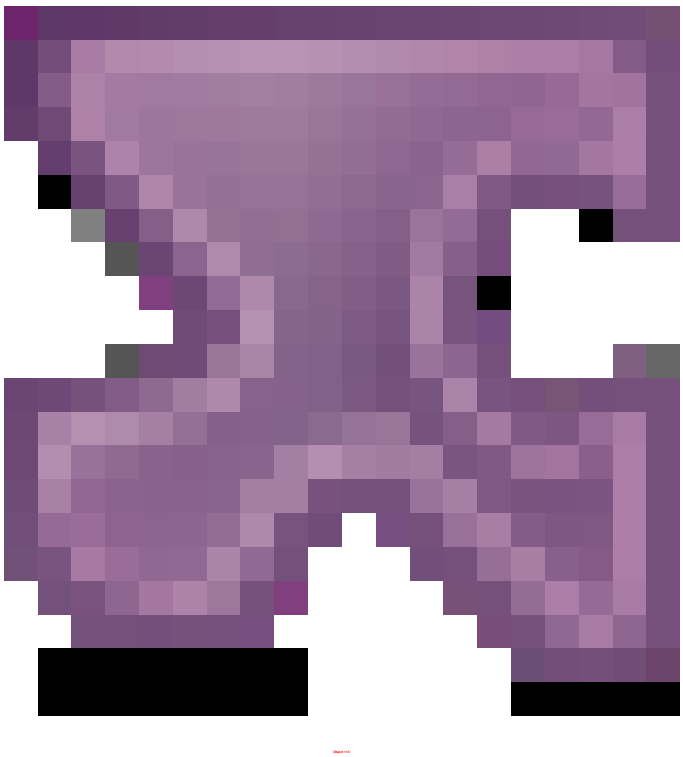

Рисунок 37 – Среднее значение невязки отдела *Spirochaetes*.

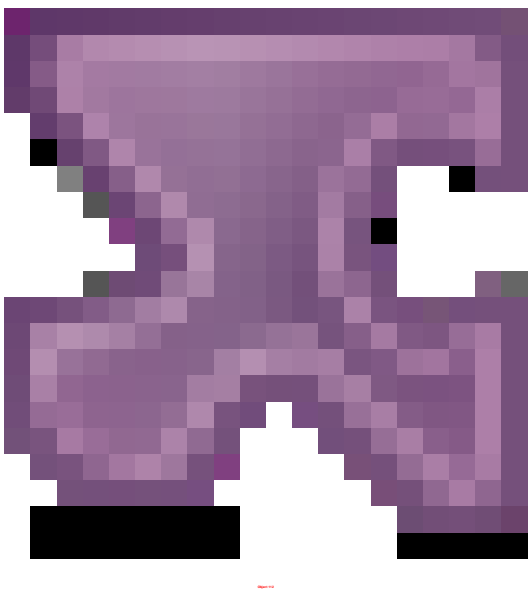

Рисунок 38 – Значение невязки отдела *Fusobacteria*, вид *Ilyobacter polytropus*.

Рисунок 39 – Среднее значение невязки отдела *Verrucomicrobia*.

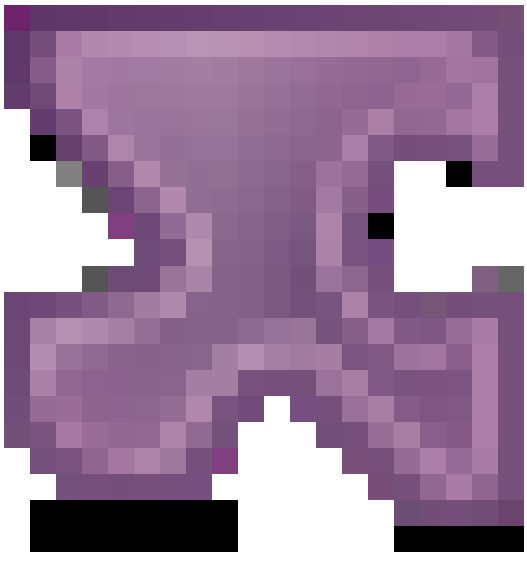

Рисунок 40 – Значение невязки отдела *Dictyoglomi*, вид *Dictyoglomus turgidum*.

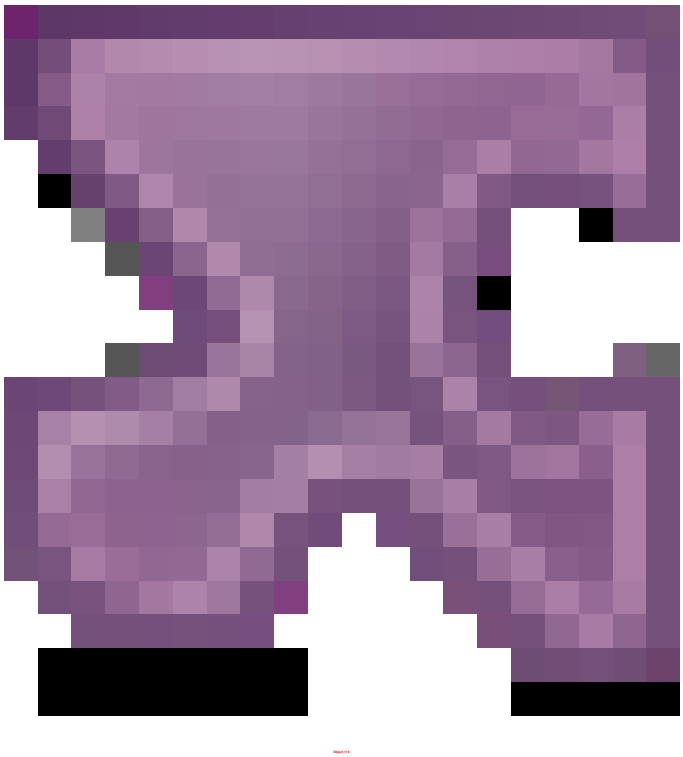

Рисунок 41 – Среднее значение невязки отдела *Actinobacteria*.

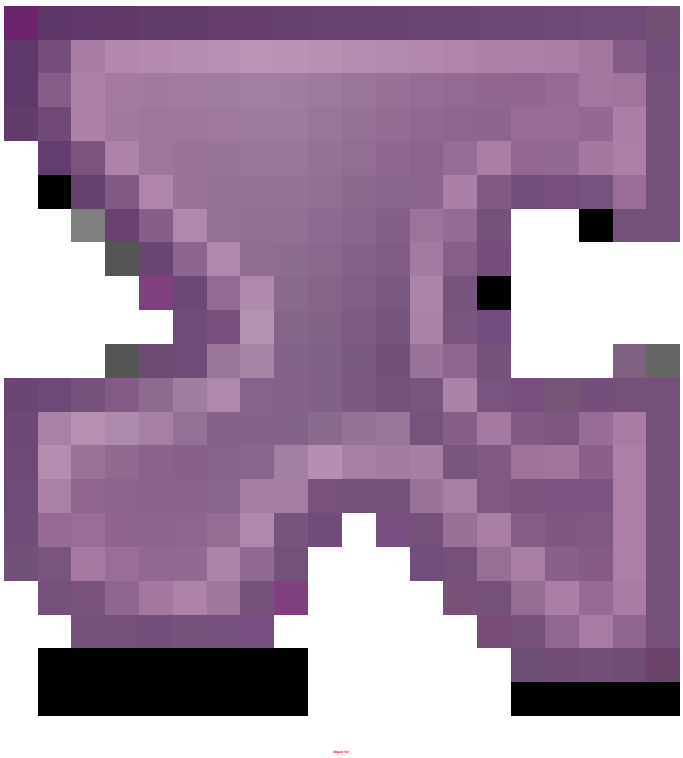

Рисунок 42 – Среднее значение невязки отдела *Synergistetes*.

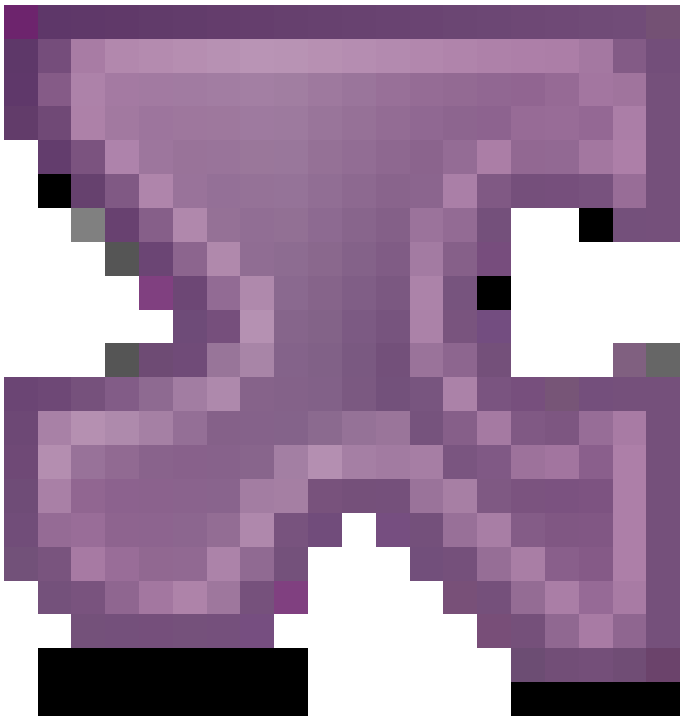

Рисунок 43 – Среднее значение невязки отдела *Tenericutes*, класс *Mollicutes*.

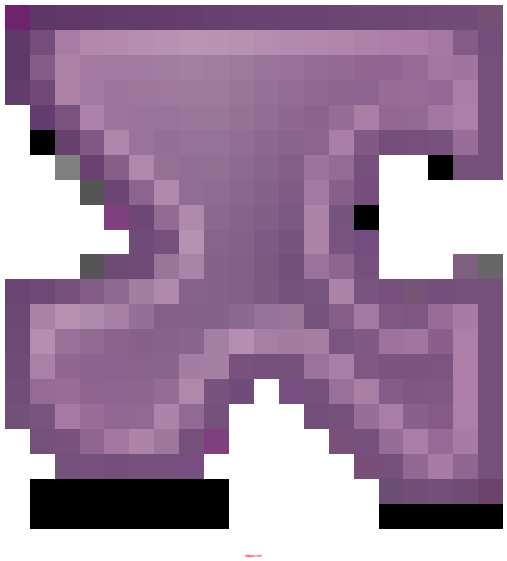

Рисунок 44 – Значение невязки отдела *Acidobacteria*, вид *Koribacter versatilis*.

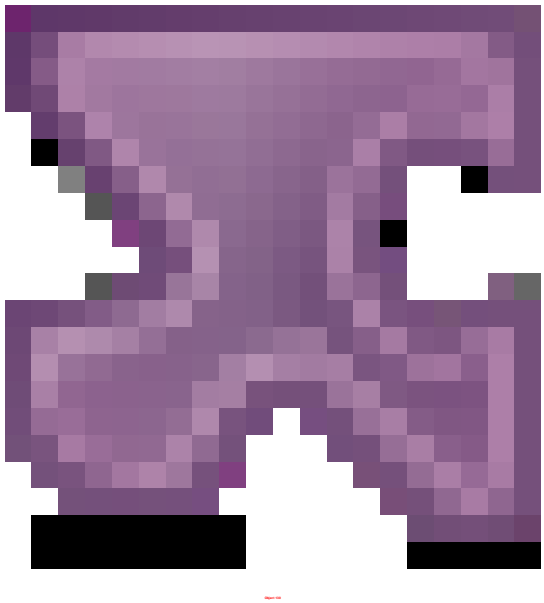

Рисунок 45 – Значение невязки отдела *Ascomycota*.

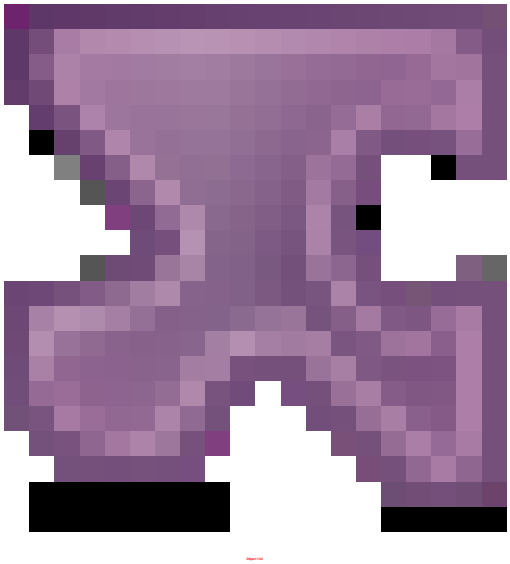

Рисунок 46 – Значение невязки отдела *Basidiomycota*, вид *Cryptococcus neoformans*.

Рисунок 47 – Значение невязки отдела *Microsporidia*, порядок *Microsporida*.

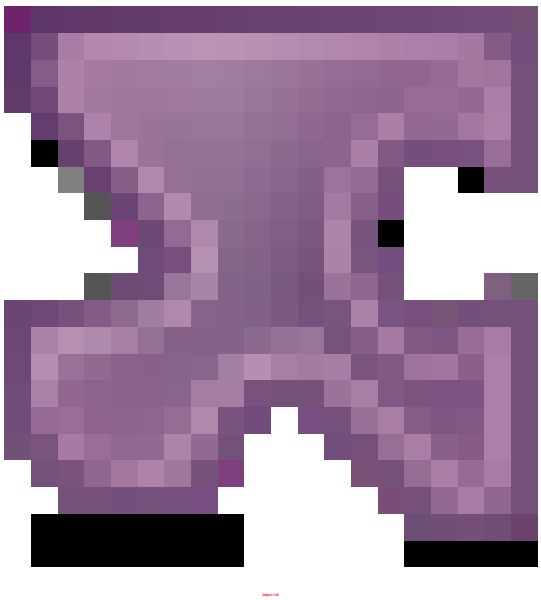

Рисунок 48 – Значение невязки отдела *Arthropoda*.

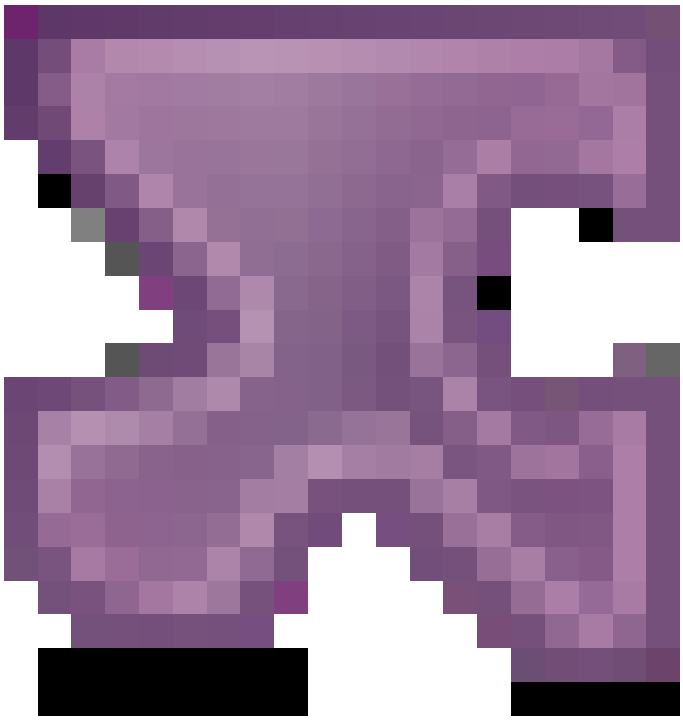

Рисунок 49 – Среднее значение невязки отдела *Chordata*.

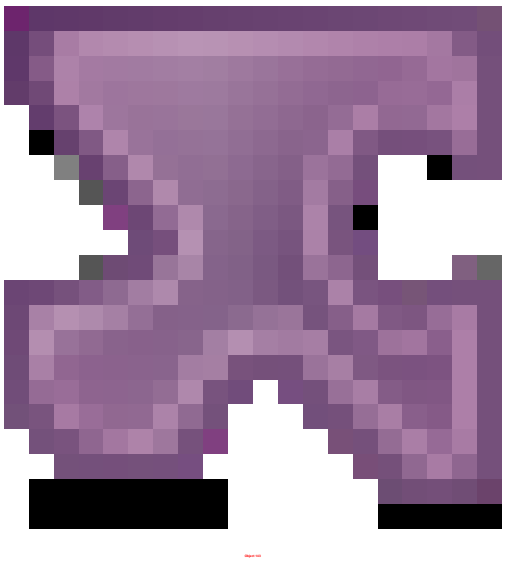

Рисунок 50 – Среднее значение невязки отдела *Nematoda*, порядок *Rhabditida*.

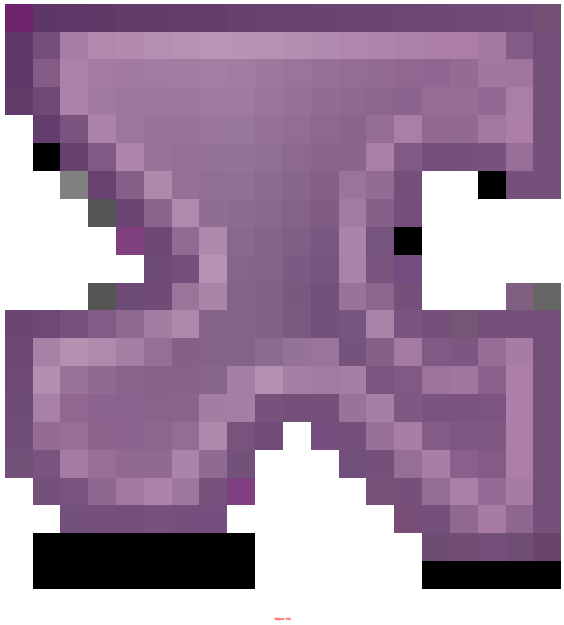

Рисунок 51 – Значение невязки отдела *Platyhelminthes*, вид *Schistosoma mansoni*.

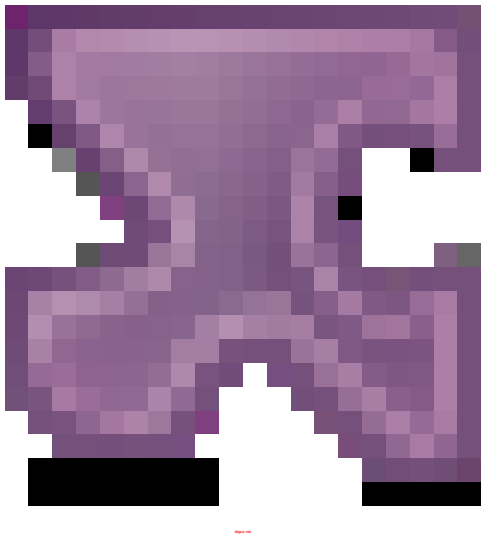

Рисунок 52 – Значение невязки отдела *Rhodophyta*, вид *Cyanidioschyzon merolae*.

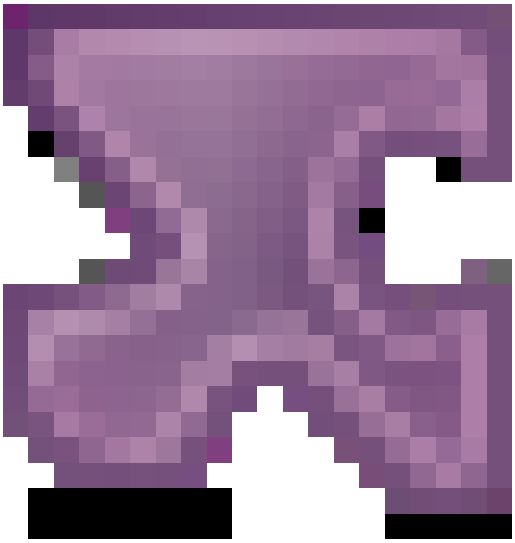

Рисунок 53 – Значение невязки отдела *Amoebozoa*, вид *Dictyostelium discoideum*.

Рисунок 54 – Значение невязки отдела *Apicomplexa*.

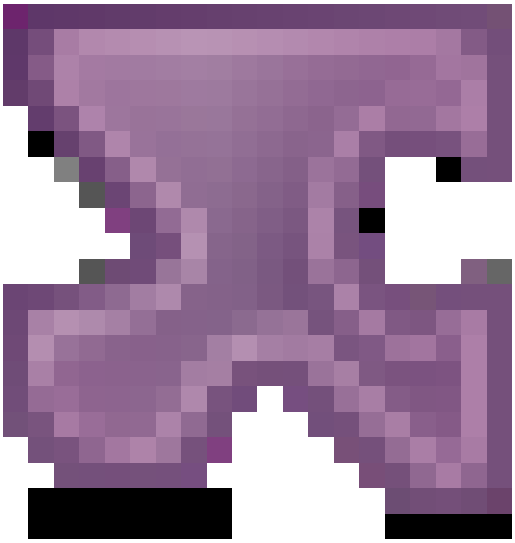

Рисунок 55 – Значение невязки отдела *Euglenozoa*, порядок *Trypanosomatidae*.

Рисунок 56 – Значение невязок отдела *Heterokontophyta*.

Рисунок 57 – Значение невязки отдела *Chlorophyta*, вид *Ostreococcus lucimarinus*.

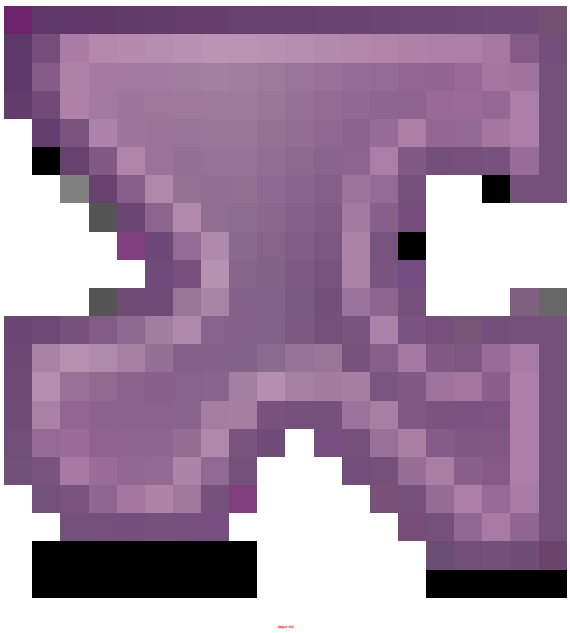

Рисунок 58 – Значение невязки отдела *Magnoliophyta*.

Так же была посчитана корреляция между значениями невязок разных порядков внутри отдела и между значениями невязок разных порядков разных отделов. Внутри отдела корреляция выше, чем между разными отделами.

Рисунок 59 – Корреляция между невязками внутри порядков разных отделов, и невязками между разными отделами архей.

Рисунок 60 – Корреляция между невязками внутри порядков разных отделов, и невязками между разными отделами бактерий.

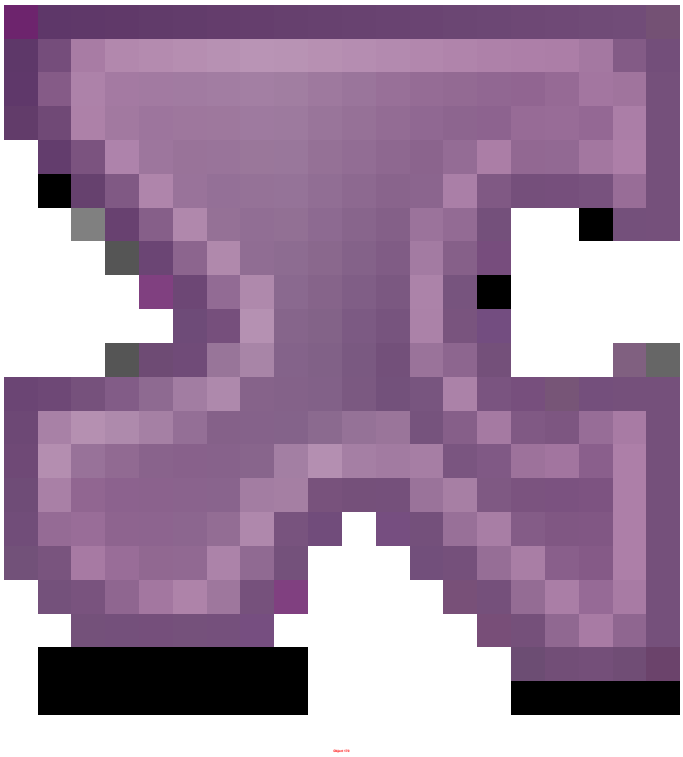

Рисунок 61 – Корреляция между невязками внутри порядков разных отделов, и невязками между разными отделами архей.

4. ВЫВОДЫ

1) Выявлена зависимость между коэффициентом невязки и таксономическим положением организмов. У организмов одного рода коэффициенты невязки принимают близкие значения. Корреляция между невязками разных порядков одного отдела выше, чем между невязками разных отделов.
2) Значение невязки падает экспоненциально с ростом толщины словаря, для любых организмов.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Усанов Н.Н., Усанов Н.Г. Симметричные структуры в комплементарных цепях геномных ДНК и возможные алгоритмы их организации. Москва : б.н., 2011 г.
2. Капун Е. Д. Разработка метода сравнения нуклеотидных последовательностей путем разбиения на фрагменты. Санкт-Петербург : б.н., 2010 г.
3. Патрушев Л. И., Минкевич И. Г. Проблема размера геномов эукариот. // Успехи биологической химии – 2007 – Том 47, С. 293–370.
4. Koonin E. V., Wolf Y. I. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world (англ.) // Nucleic acids research. — 2008. — Vol. 36, no. 21. — P. 6688-6719.
5. Atsushi Nakabachi, Atsushi Yamashita, Hidehiro Toh, Hajime Ishikawa, Helen E. Dunbar, Nancy A. Moran, Masahira Hattori. The 160-Kilobase Genome of the Bacterial Endosymbiont Carsonella // Science. 2006. V. 314. P. 267.
6. Aleksey Zimin, Kristian A. Stevens, Marc W. Crepeau, Ann Holtz-Morris, Maxim Koriabine, Guillaume Marçais, Daniela Puiu, Michael Roberts, Jill L. Wegrzyn, Pieter J. de Jong, David B. Neale, Steven L. Salzberg, James A. Yorke, Charles H. Langley. Sequencing and Assembly of the 22-Gb Loblolly Pine Genome // GENETICS March 1, 2014 V. 196 no. 3 P. 875-890
7. Бугаенко Н.Н., Горбань А.Н., Садовский М.Г. Определение информационной емкости нуклеотидных последовательностей. Красноярск: б.н., 1997 г.

8. Садовский М.Г. Об информационной емкости символьных последовательностей. // Вычислительные технологии – 2005 - Том 10, № 4 – С. 82- 90.
9. Albrecht-Bühler G. Inversions and inverted transpositions as the basis for an almost universal “format” of genome sequences // Genomics, 2008, vol. 90, pp. 297–305.
10. D. Mitchell, R. Bridge A test of Chargaff's second rule // Biochem. Biophys. Res. Commun., 340 (2006), pp. 90–94
11. Зайцева Н.А. Нарушение второго правила Чаргаффа у митохондриальных геномов и его связь с таксономией носителя. Красноярск, СФУ.
12. Гребнев Я.В., Садовский М.Г. Второе правило чаргаффа и симметрия геномов // Фундаментальные исследования – 2014, № 12 – С. 965-968.